DOCUMENT RESUME

ED 458 218 TM 033 152

AUTHOR Stemler, Steve

TITLE An Introduction to Content Analysis. ERIC Digest.

INSTITUTION ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO EDO-TM-01-06
PUB DATE 2001-06-00
NOTE 4p.
CONTRACT ED-99-CO-0032
AVAILABLE FROM ERIC Clearinghouse on Assessment and Evaluation, 1129 Shriver Laboratory, University of Maryland, College Park, MD 20742. Tel: 800-464-3742 (Toll Free).
PUB TYPE ERIC Publications (071) -- ERIC Digests in Full Text (073)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Coding; *Content Analysis; Data Collection; *Research Methodology
IDENTIFIERS ERIC Digests

ABSTRACT

This digest provides an introduction to content analysis, a systematic, replicable technique for compressing many words of a text into fewer content categories based on explicit rules of coding. This allows inferences to be made that can then be corroborated using other methods of data collection. In a technical sense, content analysis is not restricted to the domain of text, but in order to allow for replication, the technique can only be applied to data that are durable in nature. Content analysis can be a powerful tool for determining authorship, and it is useful for examining trends and patterns in documents. Content analysis also can provide an empirical basis for monitoring shifts in public opinion. K. Krippendorff has identified six questions regarding content analysis that must be addressed to clarify the data being analyzed and the target of the inferences to be drawn. Content analysis is more than a word-frequency count, although such counts may be a useful starting point. Content analysis is particularly rich because of its reliance on coding and categorizing data. Ways of defining coding units are discussed. There are also issues of reliability and validity that must be addressed in content analysis, but when used properly, it is a powerful data reduction technique. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 218

An Introduction to Content Analysis.

ERIC DIGEST

EDO-TM-01-06
June 2001

By Steve Stemler
Boston College

ERIC Clearinghouse on Assessment and Evaluation

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
❑ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 033 152

BEST COPY AVAILABLE

An Introduction to Content Analysis

Steve Stemler
Boston College

Content analysis is a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding (Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990). It allows inferences to be made which can then be corroborated using other methods of data collection (Krippendorff, 1980). Content analysis enables researchers to sift through large volumes of data with relative ease in a systematic fashion (GAO, 1996). Krippendorff (1980) notes that "[m]uch content analysis research is motivated by the search for techniques to infer from symbolic data what would be either too costly, no longer possible, or too obtrusive by the use of other techniques" (p. 51). Further, it is a useful technique for allowing us to discover and describe the focus of individual, group, institutional, or social attention (Weber, 1990). While technically content analysis is not restricted to the domain of text, in order to allow for replication, the technique can only be applied to data that are durable in nature.

Practical Applications of Content Analysis

Content analysis can be a powerful tool for determining authorship. For instance, one technique for determining authorship is to compile a list of suspected authors, examine their prior writings, and correlate the frequency of nouns or function words to help build a case for the probability of each person's authorship of the data of interest. A Bayesian technique based on word frequency was used to show that Madison was indeed the author of the Federalist papers; recently, a more holistic approach was used to determine the identity of the anonymous author of the 1992 book *Primary Colors*.

Content analysis is also useful for examining trends and patterns in documents. For example, Stemler and Bebell (1998) conducted a content analysis of school mission statements to make some inferences about what schools hold as their primary reasons for existence. One of the major research questions was whether the criteria being used to measure program effectiveness (e.g., academic test scores) were aligned with the overall program objectives or reason for existence.

Additionally, content analysis provides an empirical basis for monitoring shifts in public opinion. Data collected from the mission statements project in the late 1990s can be objectively compared to data collected at some point in the future to determine if policy changes related to standards-based reform have manifested themselves in school mission statements.

Conducting a Content Analysis

According to Krippendorff (1980), six questions must be addressed in every content analysis:

1) Which data are analyzed?
2) How are they defined?
3) What is the population from which they are drawn?
4) What is the context relative to which the data are analyzed?
5) What are the boundaries of the analysis?
6) What is the target of the inferences?

At least three problems can occur when documents are being assembled for content analysis. First, when a substantial number of documents from the population are missing, the content analysis must be abandoned. Second, inappropriate records (e.g., ones that do not match the definition of the document required for analysis) should be discarded, but a record should be kept of the reasons. Finally, some documents might match the requirements for analysis but just be uncodable because they contain missing passages or ambiguous content (GAO, 1996).

Analyzing the Data

Perhaps the most common notion in qualitative research is that a content analysis simply means doing a word-frequency count. The assumption made is that the words that are mentioned most often are the words that reflect the greatest concerns. While this may be true in some cases, there are several counterpoints to consider when using simple word frequency counts to make inferences about matters of importance.

One thing to consider is that synonyms may be used for stylistic reasons throughout a document and thus may lead the researchers to underestimate the importance of a concept (Weber, 1990). Also bear in mind that each word may not represent a category equally well. Unfortunately, there are no well-developed weighting procedures, so for now, using word counts requires the researcher to be aware of this limitation. Furthermore, Weber reminds us that, "not all issues are equally difficult to raise. In contemporary America it may well be easier for political parties to address economic issues such as trade and deficits than the history and current plight of Native American living precariously on reservations" (1990, p. 73). Finally, in performing word frequency counts, one should bear in mind that some words may have multiple meanings. For instance the word "state" could mean a political body, a situation, or a verb meaning "to speak."

A good rule of thumb to follow in the analysis is to use word frequency counts to identify words of potential interest, and then to use a Key Word In Context (KWIC) search to test for the consistency of usage of words. Most qualitative research software (e.g., NUD*IST, HyperRESEARCH; see further information at the end of this Digest) allows the researcher to pull up the sentence in which that word was used so that he or she can see the word in some context. This procedure will help to strengthen the validity of the inferences that are being made from the data.

Content analysis extends far beyond simple word counts, however. What makes the technique particularly rich and meaningful is its reliance on coding and categorizing of the data. The basics of categorizing can be summed up in these quotes: "A category is a group of words with similar meaning or connotations" (Weber, 1990, p. 37). "Categories must be mutually exclusive and exhaustive" (GAO, 1996, p. 20). Mutually exclusive categories exist when no unit falls between two data points, and each unit is represented by only one data point. The requirement of exhaustive categories is met when the data language represents all recording units without exception.

Emergent vs. *a priori* coding. There are two approaches to coding data that operate with slightly different rules. With *emergent coding*, categories are established following some preliminary examination of the data. The steps to follow are outlined in Haney, Russell, Gulek, & Fierros (1998) and will be summarized here. First, two people independently review the material and come up with a set of features that form a checklist. Second, the researchers compare notes and reconcile any differences that show up on their initial checklists. Third, the researchers use a consolidated checklist to independently apply coding. Fourth, the researchers check the reliability of the coding (a 95% agreement is suggested; .8 for Cohen's kappa). If the level of reliability is not acceptable, then the researchers repeat the previous steps. Once the reliability has been established, the coding is applied on a large-scale basis. The final stage is a periodic quality control check.

When dealing with *a priori* coding, the categories are established prior to the analysis based upon some theory. Professional colleagues agree on the categories, and the coding is applied to the data. Revisions are made as necessary, and the categories are tightened up to the point that maximizes mutual exclusivity and exhaustiveness (Weber, 1990).

Coding units. There are several different ways of defining coding units. The first way is to define them physically in terms of their natural or intuitive borders. The second way to define the recording units syntactically, that is, to use the separations created by the author, such as words, sentences, or paragraphs. A third way to define them is to use referential units. Referential units refer to the way a unit is represented. A fourth method of defining coding units is by using propositional units. Propositional units are perhaps the most complex method of defining coding units because they work by breaking down the text in order to examine underlying assumptions.

Typically, three kinds of units are employed in content analysis: sampling units, context units, and recording units.

Sampling units will vary depending on how the researcher makes meaning; they could be words, sentences, or paragraphs. In the mission statements project, the sampling unit was the mission statement.

Context units neither need be independent or separately describable. They may overlap and contain many recording units. Context units do, however, set physical limits on what kind of data you are trying to record. In the mission statements project, the context units were sentences. This was an arbitrary decision, and the context unit could just as easily have been paragraphs or entire statements of purpose.

Recording units, by contrast, are rarely defined in terms of physical boundaries. In the mission statements project, the recording unit was the idea(s) regarding the purpose of school found in the mission statements (e.g., develop responsible citizens or promote student self-worth). Thus a sentence that reads "The mission of Jason Lee school is to enhance students' social skills, develop responsible citizens, and foster emotional growth" could be coded in three separate recording units, with each idea belonging to only one category (Krippendorff, 1980).

Reliability. Weber (1990) notes: "To make valid inferences from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way" (p. 12). As Weber further notes, "reliability problems usually grow out of the ambiguity of word meanings, category definitions, or other coding rules" (p. 15). Yet, it is important to recognize that the people who have developed the coding scheme have often been working so closely on the project that they have established shared and hidden meanings of the coding. The obvious result is that the reliability coefficient they report is artificially inflated (Krippendorff, 1980). In order to avoid this, one of the most critical steps in content analysis involves developing a set of explicit recording instructions. These instructions then allow outside coders to be trained until reliability requirements are met.

Reliability may be discussed in the following terms: *Stability*, or intra-rater reliability. Can the same coder get the same results try after try? *Reproducibility*, or inter-rater reliability. Do coding schemes lead to the same text being coded in the same category by different people? One way to measure reliability is to measure the percent of agreement between raters. Another is the kappa statistic.

Validity. It is important to recognize that a methodology is always employed in the service of a research question. As such, validation of the inferences made on the basis of data from one analytic approach demands the use of multiple sources of information. If at all possible, the researcher should try to have some sort of validation study built into the design. In qualitative research, validation takes the form of triangulation. Triangulation lends credibility to the findings by incorporating multiple sources of data, methods, investigators, or theories.

Conclusion

When used properly, content analysis is a powerful data reduction technique. Its major benefit comes from the fact that it is a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding. It has the attractive features of being unobtrusive, and being useful in dealing with large volumes of data. The technique of content analysis extends far beyond simple word frequency counts. Many limitations of word counts have been discussed and methods of extending content analysis to enhance the utility of the analysis have been addressed. Two fatal flaws that destroy the utility of a content analysis are faulty definitions of categories and non-mutually exclusive and exhaustive categories.

Further information

For links, articles, software and resources see
http://writing.colostate.edu/references/research/content/
http://www.gsu.edu/~wwwcom/.

This digest is based on an article originally appearing in *Practical Assessment Research and Evaluation*

References

Berelson, B. (1952). *Content Analysis in Communication Research*. Glencoe, Ill: Free Press.

Haney, W., Russell, M., Gulek, C., and Fierros, E. (Jan-Feb, 1998). Drawing on education: Using student drawings to promote middle school improvement. *Schools in the Middle, 7(3)*, 38- 43.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

Stemler, S., and Bebell, D. (1998). *An Empirical Approach to Understanding and Analyzing the Mission Statements of Selected Educational Institutions*. Paper presented at the annual meeting of the New England Educational Research Organization. Portsmouth, New Hampshire. Available: ERIC Doc No. ED 442 202

U.S. General Accounting Office (1996). *Content Analysis: A Methodology for Structuring and Analyzing Written Material*. GAO/PEMD-10.3.1. Washington, D.C. (This book can be ordered free from the GAO).

Weber, R. P. (1990). *Basic Content Analysis*, 2nd ed. Newbury Park, CA.

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract ED99CO0032. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI or the U.S. Department of Education.

Permission is granted to copy and distribute this ERIC/AE Digest.